

A Markov chain Monte Carlo method for estimating the statistical significance of proteoform identifications by top-down mass spectrometry (Supplementary Material)

Qiang Kou¹, Zhe Wang², Rachele A. Lubeckyj³, Si Wu², Liangliang Sun³, and Xiaowen Liu^{1,4,*}

¹Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis

²Department of Chemistry and Biochemistry, The University of Oklahoma

³Department of Chemistry, Michigan State University

⁴Center for Computational Biology and Bioinformatics, Indiana University School of Medicine

Contents

S1 Estimation of p-values of PrSMs	S3
S2 The eTDA method for FDR estimation	S3

List of Figures

Figure S1 Histogram of the numbers of cousin proteins for 33 000 random proteins with a molecular mass similar to 12454.60 Dalton.	S5
Figure S2 The D values between the distributions of the p -values reported by TopMCMC for the 2 638 entrapment PrSMs from the histone H3 data set and the uniform distribution over $[0, 1]$ with various settings of the parameter c_{max} , the number of simulations.	S5
Figure S3 The D values between the distributions of the p -values reported by TopMCMC for the 2 638 entrapment PrSMs from the histone H3 data set and the uniform distribution over $[0, 1]$ with various settings of the parameter T	S6
Figure S4 A proteoform of the protein S100-A11 (UniProt ID: P31949) with two acetylation sites and two dimethylation sites.	S6
Figure S5 A proteoform of the Acyl-CoA-binding protein (UniProt ID: P07108) with two acetylation sites and two dimethylation sites.	S7
Figure S6 A proteoform of the 10 kDa heat shock protein (UniProt ID: P61604) with two acetylation sites and two dimethylation sites.	S7
Figure S7 A proteoform of the Thymosin beta-4 protein (UniProt ID: P62328) with two acetylation sites and two dimethylation sites.	S8

List of Tables

Table S1 Five variable PTMs used in the identification of histone proteoforms	S9
Table S2 Parameter settings of TopMG in the analysis of the histone H4 data set	S9
Table S4 Parameter settings of TopMG in the analysis of the histone H3 data set against the bipartite database	S9

Table S5	Parameter settings of TopMCMC in the analysis of the histone H4 data set	S9
Table S7	Parameter settings of TopPIC in the analysis of the EC data set	S10
Table S9	Parameter settings of TopMG in the analysis of the EC data set	S10
Table S10	Parameter settings of TopPIC in the analysis of the MCF-7 data set	S11
Table S12	Parameter settings of TopMG in the analysis of the MCF-7 data set	S11

Tables in Excel files

Table S3	A total of 1 112 PrSMs were identified from the histone H4 data set by TopMG.
Table S6	A total of 2 638 entrapment PrSMs were identified from the histone H3 data set by TopMG.
Table S8	A total of 1 920 PrSMs were identified from the EC data set by TopPIC.
Table S11	A total of 615 PrSMs were identified from the MCF-7 data set by TopPIC.
Table S13	A total of 161 PrSMs were identified from the MCF-7 data set by TopMCMC.

Supplementary files

File S1	A total of 161 PrSMs were identified from the MCF-7 data set by TopMCMC.
---------	--

S1 Estimation of p -values of PrSMs

Given a spectrum S , a random protein set D , and a set of multisets Φ_k , we estimate the p -value of a PrSM with a score t . A protein subset $D_{i,j}$ of D is matchable if the difference between the residue mass of S and the mass j is explained by a multiset $V \in \Phi_k$. Let D^* be the set of all matchable subsets $D_{i,j}$. If a protein is not in a matchable subset, then its P-score is always zero and can be ignored in the computation of p -values. As a result,

$$\max_{P \in D, V \in \Phi_k} \text{Score}(S, P, V) = \max_{D_{i,j} \in D^*} \max_{P \in D_{i,j}} \max_{V \in \Phi_k} \text{Score}(S, P, V).$$

The expected value of $Y(k, t)$ is

$$\begin{aligned} E(Y(k, t)) &= \sum_{V \in \Phi_k} E(X(t, V)) \\ &= \sum_{V \in \Phi_k} \sum_i \sum_j p(i, j, t, V) \cdot d_{i,j} \\ &= \sum_i \sum_j \sum_{V \in \Phi_k} p(i, j, t, V) \cdot d_{i,j} \\ &= \sum_{D_{i,j} \in D^*} \sum_{V \in \Phi_k} p(i, j, t, V) \cdot d_{i,j} \\ &= \sum_{D_{i,j} \in D^*} \sum_{P \in D_{i,j}} \sum_{V \in \Phi_k} p(i, j, t, V). \end{aligned}$$

It is common that the difference between the residue masses of spectrum S and protein $P \in D_{i,j}$ is explained by only one multiset $V \in \Phi_k$. In this case, the probability $p(i, j, v, V)$ is non-zero for only the multiset V . We estimate the probability

$$\Pr(\max_{V \in \Phi_k} \text{Score}(S, P, V) \geq t)$$

as

$$\sum_{V \in \Phi_k} \Pr(\text{Score}(S, P, V) \geq t) = \sum_{V \in \Phi_k} p(i, j, t, V).$$

Let $e = E(Y(k, t))$ and z be the number of all matchable proteins. That is, $z = \sum_{D_{i,j} \in D^*} |D_{i,j}|$. The value $\sum_{V \in \Phi_k} p(i, j, t, V)$ can be estimated as e/z . That is,

$$\Pr(\max_{V \in \Phi_k} \text{Score}(S, P, V) \geq t) \approx \frac{e}{z}.$$

We assume that the scores for all proteins are independent. The p -value is the probability that $\max_{P \in D, V \in \Phi_k} \text{Score}(S, P, V) \geq t$, which is estimated by $1 - (1 - e/z)^z$.

S2 The eTDA method for FDR estimation

Here we give a brief description of the eTDA estimator framework. Please refer to Ref [1] for more details.

In the eTDA estimator framework, any scoring function can be used. Given a spectrum σ , a set of variable PTMs V , a database DB , we use $\text{Score}(\sigma, DB, V)$ to represent the E -value of the best PrSM between σ and a proteoform of a protein in DB that may contain variable PTMs in V . When the set V is fixed, we use $\text{Score}(\sigma, DB)$ for $\text{Score}(\sigma, DB, V)$.

Given a cutoff t , an identification is reported in DB if $\text{Score}(\sigma, DB) \leq t$. The DB can be a target database T , a decoy database R , or a concatenated database $T \oplus R$. In a target-decoy search, a decoy

identification is reported if $\text{Score}(\sigma, R) \leq \min(t, \text{Score}(\sigma, T))$. This means the decoy identification has a better score in the decoy database than the target database, and the score of the decoy identification is better than the cutoff t . The total number of decoy identifications in the spectrum set Σ is computed as

$$DD(\Sigma, T \oplus R, t) = \sum_{\sigma \in \Sigma} 1_{\text{Score}(\sigma, R) \leq \min(t, \text{Score}(\sigma, T))}$$

where the random variable 1_A is equal to 1 if and only if the event A occurred.

Similarly, the total number of target identifications is computed as

$$TD(\Sigma, T \oplus R, t) = \sum_{\sigma \in \Sigma} 1_{\text{Score}(\sigma, T) \leq \min(t, \text{Score}(\sigma, R))}$$

Then the false discovery rate (FDR) is estimated as

$$\widehat{FDR}_{TDA} = \frac{2 \cdot DD(\Sigma, T \oplus R, t)}{DD(\Sigma, T \oplus R, t) + TD(\Sigma, T \oplus R, t)}$$

If the size of the spectra set Σ is large enough, we have

$$E[\widehat{FDR}_{TDA}] = E\left[\frac{2 \cdot DD(\Sigma, T \oplus R, t)}{DD(\Sigma, T \oplus R, t) + TD(\Sigma, T \oplus R, t)}\right] \approx \frac{2 \cdot E[DD(\Sigma, T \oplus R, t)]}{E[DD(\Sigma, T \oplus R, t)] + E[TD(\Sigma, T \oplus R, t)]}$$

This is defined as the *eTDA estimator*:

$$\widehat{FDR}_{eTDA} := \frac{2 \cdot E[DD(\Sigma, T \oplus R, t)]}{E[DD(\Sigma, T \oplus R, t)] + E[TD(\Sigma, T \oplus R, t)]}$$

$E[DD(\Sigma, T \oplus R, t)]$ is calculated as

$$\begin{aligned} E[DD(\Sigma, T \oplus R, t)] &= E\left[\sum_{\sigma \in \Sigma} 1_{\text{Score}(\sigma, R) \leq \min(t, \text{Score}(\sigma, T))}\right] \\ &= \sum_{\sigma \in \Sigma} E[1_{\text{Score}(\sigma, R) \leq \min(t, \text{Score}(\sigma, T))}] \\ &= \sum_{\sigma \in \Sigma} P[\text{Score}(\sigma, R) \leq \min(t, \text{Score}(\sigma, T))] \end{aligned}$$

Similarly, $E[TD(\Sigma, T \oplus R, t)]$ is calculated as

$$\begin{aligned} E[TD(\Sigma, T \oplus R, t)] &= E\left[\sum_{\sigma \in \Sigma} 1_{\text{Score}(\sigma, T) \leq \min(t, \text{Score}(\sigma, R))}\right] \\ &= \sum_{\sigma \in \Sigma} E[1_{\text{Score}(\sigma, T) \leq \min(t, \text{Score}(\sigma, R))}] \\ &= \sum_{\sigma \in \Sigma} P[\text{Score}(\sigma, T) \leq \min(t, \text{Score}(\sigma, R))] \\ &= \sum_{\sigma \in \Sigma: \text{Score}(\sigma, T) \leq t} P[\text{Score}(\sigma, T) < \text{Score}(\sigma, R)] \\ &= \sum_{\sigma \in \Sigma: \text{Score}(\sigma, T) \leq t} \left(1 - P[\text{Score}(\sigma, R) < \text{Score}(\sigma, T)]\right) \end{aligned}$$

When a cutoff value t' is very small, the probability $P[\text{Score}(\sigma, R) \leq t']$ is computed as

$$P[\text{Score}(\sigma, R) \leq t'] \approx t'.$$

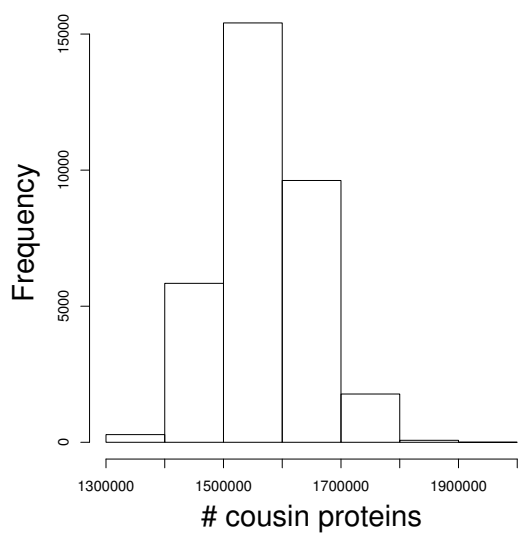


Figure S1: Histogram of the numbers of cousin proteins for 33 000 random proteins with a molecular mass similar to 12454.60 Dalton.

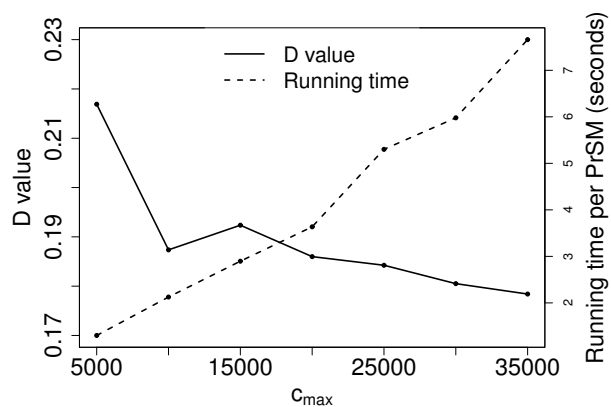


Figure S2: The D values between the distributions of the p -values reported by TopMCMC for the 2638 entrapment PrSMs from the histone H3 data set and the uniform distribution over $[0, 1]$ with various settings of the parameter c_{max} , the number of simulations.

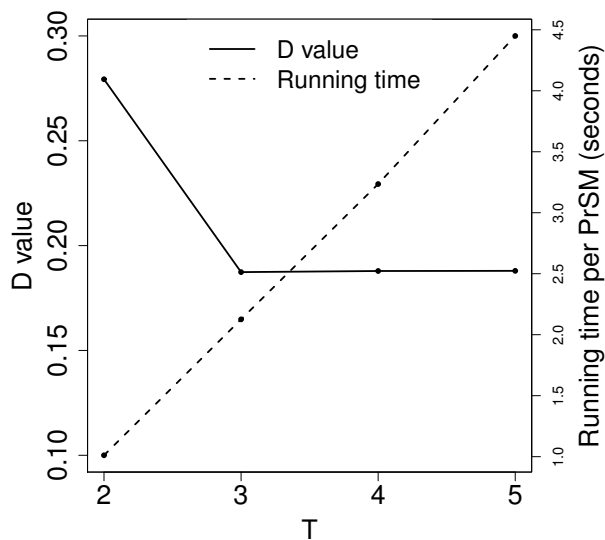


Figure S3: The D values between the distributions of the p -values reported by TopMCMC for the 2638 entrapment PrSMs from the histone H3 data set and the uniform distribution over $[0, 1]$ with various settings of the parameter T .

All proteins / sp|P31949|S10AB_HUMAN Protein S100-A11 OS=Homo sapiens OX=9606 GN=S100A11 PE=1 SV=2 / Proteoform #8

Protein-Spectrum-Match #22 for Spectrum #404

PrSM ID:	22	Scan(s):	1288	Precursor charge:	13
Precursor m/z:	913.0011	Precursor mass:	11855.9199	Proteoform mass:	11855.8948
# matched peaks:	55	# matched fragment ions:	41	# unexpected modifications:	0
E-value:	8.24e-08	P-value:	1.03e-06	Q-value (Spectral FDR):	0

Acetyl

1 M]A K I S S P T E T E R C I E S L I A V } F Q K Y A G K D G Y 30

31 N Y T L S K T E F L S F M N T E L A A F T K N Q K D P G V L 60

61 D R M M K K L D T N S D G Q L D F S E F L N L I G L A M A 90

91 C H D S F L K A V P S Q K R T 105

Fixed PTMs: Carbamidomethylation [C13 C91]

Variable PTMs: Acetyl [A2 - L17] Dimethyl;Dimethyl;Acetyl [P57 - K66]

All peaks (197) Matched peaks (55) Not matched peaks (142)

Figure S4: A proteoform of the protein S100-A11 (UniProt ID: P31949) with two acetylation sites and two dimethylation sites.

Protein-Spectrum-Match #26 for Spectrum #417

PrSM ID:	26	Scan(s):	1305	Precursor charge:	13
Precursor m/z:	773.8601	Precursor mass:	10047.0865	Proteoform mass:	10047.2465
# matched peaks:	43	# matched fragment ions:	34	# unexpected modifications:	0
E-value:	2.37e-08	P-value:	7.90e-07	Q-value (Spectral FDR):	0

Dimethyl;Acetyl
 1 M] S Q A E F E K A A E] E V R] H L K T K P S D E E M L] F I I Y L G 30
 31 L H Y K Q A T V L G D I L N T L E R P G M L D F T G K A K L W D L A W N 60
 61 E L] K L G T S K E L D L A M K L A L Y I L N K L V L E L E L L K L K K Y G I 87

Variable PTMs: Dimethyl;Acetyl [S2 - E11] Dimethyl [E12 - R14] Acetyl [H15 - L26]

All peaks (167) Matched peaks (43) Not matched peaks (124)

Figure S5: A proteoform of the Acyl-CoA-binding protein (UniProt ID: P07108) with two acetylation sites and two dimethylation sites.

Protein-Spectrum-Match #57 for Spectrum #563

PrSM ID:	57	Scan(s):	1511	Precursor charge:	14
Precursor m/z:	782.0033	Precursor mass:	10933.9437	Proteoform mass:	10933.9387
# matched peaks:	38	# matched fragment ions:	26	# unexpected modifications:	0
E-value:	5.10e-09	P-value:	1.70e-07	Q-value (Spectral FDR):	0

Acetyl
 1 M] A G Q A F R K F L P L F D R V L V E R S A A E T V] T K G G 30
 31 I M] L P E] K] S Q G K V L] Q] A] T] V] V] A] V L G S G S K G K G G E I 60
 61 Q P V S V K V G D K V L L L P L E Y G G T K V V L D D K L D Y F L 90
 91 F R D G D I L L G K Y V D 102

Variable PTMs: Acetyl [A2 - V26] Dimethyl;Dimethyl;Acetyl [S51 - V71]

All peaks (182) Matched peaks (38) Not matched peaks (144)

Figure S6: A proteoform of the 10 kDa heat shock protein (UniProt ID: P61604) with two acetylation sites and two dimethylation sites.

Protein-Spectrum-Match #35 for Spectrum #492

PrSM ID:	35	Scan(s):	1416	Precursor charge:	6
Precursor m/z:	844.0834	Precursor mass:	5058.4565	Proteoform mass:	5058.6065
# matched peaks:	17	# matched fragment ions:	17	# unexpected modifications:	0
E-value:	6.81e-10	P-value:	6.81e-08	Q-value (Spectral FDR):	0

Acetyl
 1 M] S D K] P D] M A E] I] E] K F D] K S K L K] K T E T Q E] K] N] P L P 30
 31 S K E] T] I] E] Q] E] K] Q A G E S 44

Variable PTMs: Acetyl [S2 - K4] Dimethyl;Dimethyl;Acetyl [K15 - K19]

All peaks (84) Matched peaks (17) Not matched peaks (67)

Figure S7: A proteoform of the Thymosin beta-4 protein (UniProt ID: P62328) with two acetylation sites and two dimethylation sites.

Table S1: Five variable PTMs used in the identification of histone proteoforms

PTM	Monoisotopic mass shift (Da)	Amino acids
Acetylation	42.01056	R, K
Methylation	14.01565	R, K
Dimethylation	28.03130	R, K
Trimethylation	42.04695	R
Phosphorylation	79.96633	S, T, Y

Table S2: Parameter settings of TopMG in the analysis of the histone H4 data set

Parameter	Value
Fragmentation method	FILE
Fixed modifications	None
N-terminal forms of proteins	NONE, NME, NME+ACETYLTATION
Using a decoy database	No
Error tolerance	15 ppm
Maximum number of unexpected modifications (unknown mass shifts) in a PrSM	0
Number of combined spectra	1
Gap in constructing proteoform graph	40
Maximum number of variable modifications	10
Maximum number of variable PTMs in a graph gap	5

Table S4: Parameter settings of TopMG in the analysis of the histone H3 data set against the bipartite database

Parameter	Value
Fragmentation method	FILE
Fixed modifications	None
N-terminal forms of proteins	NONE, NME, NME+ACETYLTATION
Using a decoy database	No
Error tolerance	15 ppm
Maximum number of unexpected modifications (unknown mass shifts) in a PrSM	0
Number of combined spectra	1
Gap in constructing proteoform graph	40
Maximum number of variable modifications	5
Maximum number of variable PTMs in a graph gap	5

Table S5: Parameter settings of TopMCMC in the analysis of the histone H4 data set

Parameter	Value
T : the number of rounds for estimating oversampling factors	3
c_{max} : the number of simulations	10 000
Error tolerance	15 ppm

Table S7: Parameter settings of TopPIC in the analysis of the EC data set

Parameter	Value
Number of combined spectra	1
Fragmentation method	FILE
Search type	TARGET+DECOY
Fixed modifications	None
Use TopFD feature file:	True
Maximum number of unexpected modifications	1
Error tolerance	15 ppm
Spectrum-level cutoff type	FDR
Spectrum-level cutoff value	0.01
Proteoform-level cutoff type	FDR
Proteoform-level cutoff value	0.01
Allowed N-terminal forms	NONE, NME, NME+ACETYLATION, METHIONINE ACETYLATION
Maximum mass shift of modifications	500 Da
Minimum mass shift of modifications	−500 Da
Thread number	1
E-value computation	Lookup table

Table S9: Parameter settings of TopMG in the analysis of the EC data set

Parameter	Value
Fragmentation method	FILE
Search type	TARGET+DECOY
Fixed modifications	None
Use TopFD feature file	True
Error tolerance	15 ppm
Allowed N-terminal forms	NONE, NME, NME+ACETYLATION, METHIONINE ACETYLATION
Maximum mass shift of modifications	500 Da
Thread number:	16
Gap in proteoform graph:	40
Maximum number of variable PTMs	5
Maximum number of variable PTMs in a graph gap	5

Table S10: Parameter settings of TopPIC in the analysis of the MCF-7 data set

Parameter	Value
Number of combined spectra	1
Fragmentation method	FILE
Search type	TARGET+DECOY
Fixed modifications	C57
Use TopFD feature file:	True
Maximum number of unexpected modifications	1
Error tolerance	15 ppm
Spectrum-level cutoff type	FDR
Spectrum-level cutoff value	0.01
Proteoform-level cutoff type	FDR
Proteoform-level cutoff value	0.01
Allowed N-terminal forms	NONE, NME, NME+ACETYLATION, METHIONINE ACETYLATION
Maximum mass shift of modifications	500 Da
Minimum mass shift of modifications	−500 Da
Thread number	1
E-value computation	Lookup table

Table S12: Parameter settings of TopMG in the analysis of the MCF-7 data set

Parameter	Value
Fragmentation method	FILE
Search type	TARGET+DECOY
Fixed modifications	C57
Use TopFD feature file	True
Error tolerance	15 ppm
Allowed N-terminal forms	NONE, NME, NME+ACETYLATION, METHIONINE ACETYLATION
Maximum mass shift of modifications	500 Da
Thread number:	16
Gap in proteoform graph:	40
Maximum number of variable PTMs	5
Maximum number of variable PTMs in a graph gap	5

References

- [1] Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry*, 22:1111–1120, 2011.